(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **11250100 A**

(43) Date of publication of application: **17 . 09 . 99**

(51) Int. Cl **G06F 17/30**
G06F 15/18
G06F 17/21

(21) Application number: **10064682**

(22) Date of filing: **27 . 02 . 98**

(71) Applicant: **NEC CORP**

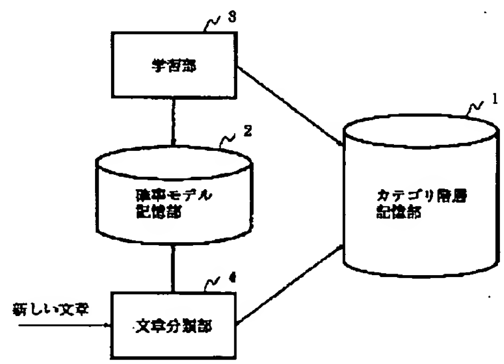
(72) Inventor: **RI KO**
YAMANISHI KENJI

(54) **HIERARCHICAL DOCUMENT CLASSIFYING DEVICE AND MACHINE-READABLE RECORDING MEDIUM RECORDING PROGRAM**

(57) Abstract:

PROBLEM TO BE SOLVED: To classify a sentence to categorical hierarchies based on the distribution of words (keyword) appearing in the sentence (text and document).

SOLUTION: This method grasps document classification as a statistical test problem. A categorical hierarchy storing part 1 stores category hierarchies. A probability model storing part 2 stores linear combination models. A learning part 3 refers to categorical hierarchies stored in the part 1, learns a linear combination model corresponding to each category from a document that is already classified to a category and stores the linear combination model in the part 2. A document classifying part 4 newly inputs a document, refers to each category in category hierarchies stored in the part 1, refers to a linear combination model corresponding to each category from the part 2 for the category, calculates the negative logarithmic likelihood of each linear combination model to the inputted document and classifies the inputted document to a category corresponding to a linear combination model that has the smallest negative logarithmic likelihood.



(19)日本国特許庁（ＪＰ）

(12) 公開特許公報（Ａ）

(11)特許出願公開番号
特開平11－250100

(43)公開日 平成11年(1999) 9月17日

(51)Int.Cl.6		識別記号	FI		
Ｇ０６Ｆ	17/30		Ｇ０６Ｆ	15/401	３１０Ｄ
	15/18	５６０		15/18	５６０Ａ
	17/21			15/20	５７０Ｚ
					５９０Ｚ
				15/40	３７０Ａ

審査請求 有　請求項の数４　ＦＤ　(全 11 頁)

(21)出願番号　特願平10－64682

(22)出願日　平成10年(1998) 2月27日

(71)出願人 000004237
日本電気株式会社
東京都港区芝五丁目７番１号

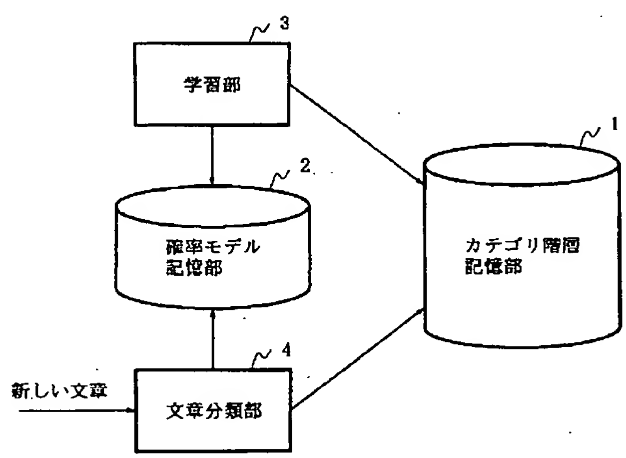
(72)発明者 李 航
東京都港区芝五丁目７番１号 日本電気株式会社内

(72)発明者 山西 健司
東京都港区芝五丁目７番１号 日本電気株式会社内

(74)代理人 弁理士 境 廣巳

(54)【発明の名称】 階層型文章分類装置およびプログラムを記録した機械読み取り可能な記録媒体

(57)【要約】
【課題】 文章（テキスト、ドキュメント）に現れる単語（キーワード）の分布を基に文章をカテゴリの階層に分類する。
【解決手段】 文章分類を統計的検定問題として捉える。カテゴリ階層記憶部１ではカテゴリの階層が記憶される。確率モデル記憶部２では、線形結合モデルが記憶される。学習部３は、カテゴリ階層記憶部１に記憶されるカテゴリの階層を参照し、既にカテゴリに分類された文章から各カテゴリの対応する線形結合モデルを学習し、線形結合モデルを確率モデル記憶部２に記憶する。文章分類部４は、新しく文章を入力し、カテゴリ階層記憶部１に記憶されるカテゴリの階層における各カテゴリを参照し、各カテゴリに対して、確率モデル記憶部２から、そのカテゴリに対応する線形結合モデルを参照し、入力文章に対する各線形結合モデルの負対数尤度を計算し、負対数尤度の最も小さい線形結合モデルに対応するカテゴリに入力文章を分類する。



【特許請求の範囲】

【請求項１】　ノードが文章の分類されたカテゴリを表現し、リンクがカテゴリの上位下位関係を表現するグラフとして、カテゴリの階層を記憶するカテゴリ階層記憶部、

前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに対して、より下位のカテゴリの単語空間上の確率モデルの重みつき平均を該カテゴリの線形結合モデルとし、各カテゴリの線形結合モデルを記憶する確率モデル記憶部、

前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに分類された文章を基に、各カテゴリの線形結合モデルを、より下位のカテゴリの線形結合モデルから学習し、学習できた各カテゴリの線形結合モデルを前記確率モデル記憶部に記憶する学習部、

新しく文章を入力し、該入力文章を単語のデータ列と見なし、前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに対して、前記確率モデル記憶部に記憶される該カテゴリの線形結合モデルの該入力文章に対する負対数尤度を計算し、計算された負対数尤度の最も小さいカテゴリに該入力文章を分類する文章分類部、

を備えることを特徴とする階層型文章分類装置。

【請求項２】　ノードが文章の分類されたカテゴリを表現し、リンクがカテゴリの上位下位関係を表現するグラフとして、カテゴリの階層を記憶するカテゴリ階層記憶部、

前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに対して、より下位のカテゴリの、単語空間上の確率モデルの集合を該カテゴリの確率モデルの集合とし、各カテゴリの確率モデルの集合の全ての要素を記憶する確率モデル集合記憶部、

前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに分類された文章を基に、各カテゴリの確率モデルの集合を、より下位のカテゴリの単語空間上の確率モデルの集合から学習し、学習できた各カテゴリの確率モデルの集合のすべての要素を前記確率モデル集合記憶部に記憶する学習部、

新しく文章を入力し、該入力文章を単語のデータ列と見なし、前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに対して、前記確率モデル集合記憶部に記憶される該カテゴリの確率モデルの集合に対する該入力文章の確率的複雑度を計算し、計算された確率的複雑度の最も小さいカテゴリに該入力文章を分類する文章分類部、

を備えることを特徴とする階層型文章分類装置。

【請求項３】　コンピュータを、請求項１に記載する、カテゴリ階層記憶部、確率モデル記憶部、学習部、および文章分類部として機能させるプログラムを記録した機械読み取り可能な記録媒体。

【請求項４】　コンピュータを、請求項２に記載する、

カテゴリ階層記憶部、確率モデル集合記憶部、学習部、および文章分類部として機能させるプログラムを記録した機械読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、インターネットのホームページの自動分類、電子図書館における文献検索、特許出願情報の検索、電子化された新聞記事の自動分類、マルチメディア情報の自動分類等の情報の分類や検索に関するものである。

【０００２】

【従来の技術】情報の分類や検索の分野では、文章分類（ドキュメント分類、テキスト分類ともいう）装置の開発は大きな課題である。ここでいう文章分類とは、予め人間がカテゴリを設け、さらに一部の文章がそれぞれどのカテゴリに属するかを判断し、該当のカテゴリにそれらの文章を分類し、システムに記憶した後、システムは記憶された情報から知識を自動的に獲得し、それ以後、獲得できた知識を基に、新たに入力された文章を自動的に分類することを指す。

【０００３】文章はカテゴリに分類されているので、文章を検索する時、関係するカテゴリにおける文章だけを検索すればよく、検索が効率良く且つ正確になる。

【０００４】従来、幾つかの文章分類装置が提案されている。中でも、Ｓａｌｔｏｎらの提案する文章分類装置が良く知られている（Ｇ．Ｓａｌｔｏｎ　ａｎｄ　Ｍ．Ｊ．ＭｃＧｉｌｌ，Ｉｎｔｒｏｄｕｃｔｉｏｎ　ｔｏ　Ｍｏｄｅｒｎ　Ｉｎｆｏｒｍａｔｉｏｎ　Ｒｅｔｒｉｅｖａｌ，Ｎｅｗ　Ｙｏｒｋ：　ＭｃＧｒａｗ　Ｈｉｌｌ，１９８３）。その文章分類装置は、文章に現れる単語の頻度ベクトルとカテゴリにおける単語の頻度ベクトルとの間のコサイン値を文章とカテゴリ間の距離と見なし、距離の最も小さいカテゴリに文章を分類することを特徴としている。

【０００５】

【発明が解決しようとする課題】しかし、従来方式のほとんどは、文章を幾つかの並列のカテゴリに分類するもので、階層構造をなすカテゴリに文章を自動的に分類する装置がなかった。例えば、「政治」のカテゴリがさらに「国会」や「政党」のサブカテゴリに分かれ、文章を「政治」のカテゴリに分類した後、さらにそれを「国会」と「政党」に分類した方が後の検索がさらに高速になる。

【０００６】本発明の目的は、並列のカテゴリに文章を分類するのではなく、階層構造をなすカテゴリに文章を自動分類し得るようにすることにある。

【０００７】また、本発明の別の目的は、信頼性の高い文章の自動分類を実現することにある。

【０００８】

【課題を解決するための手段】本発明では、カテゴリを

階層化し、各カテゴリに線形結合モデルと呼ばれる確率モデル、或いは確率モデルの集合を対応させ、新しい文章が入力されると、その文章に対する線形結合モデルの負対数尤度、或いは確率モデル集合の確率的複雑度を計算し、負対数尤度の最も小さい、或いは確率的複雑度の最も小さいカテゴリに新しい文章を分類する。

【０００９】つまり、本発明では、文章における単語の分布を基にその文章をカテゴリに分類している。特に、確率的なモデルを用いた統計的検定によって文章を分類することが特徴である。

【００１０】具体的には、本発明の第１の階層型文章分類装置は、ノードが文章の分類されたカテゴリを表現し、リンクがカテゴリの上位下位関係を表現するグラフとして、カテゴリの階層を記憶するカテゴリ階層記憶部、前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに対して、より下位のカテゴリの単語空間上の確率モデルの重みつき平均を該カテゴリの線形結合モデルとし、各カテゴリの線形結合モデルを記憶する確率モデル記憶部、前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに分類された文章を基に、各カテゴリの線形結合モデルを、より下位のカテゴリの線形結合モデルから学習し、学習できた各カテゴリの線形結合モデルを前記確率モデル記憶部に記憶する学習部、新しく文章を入力し、該入力文章を単語のデータ列と見なし、前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに対して、前記確率モデル記憶部に記憶される該カテゴリの線形結合モデルの該入力文章に対する負対数尤度を計算し、計算された負対数尤度の最も小さいカテゴリに該入力文章を分類する文章分類部、を備えることを特徴とする。

【００１１】このように構成された第１の階層型文章分類装置にあっては、学習部が、カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに例えば事前に人手によって分類された文章を基に、各カテゴリの線形結合モデルを、より下位のカテゴリの線形結合モデルから学習し、学習できた各カテゴリの線形結合モデルを確率モデル記憶部に記憶し、その後、自動分類対象となる文章が入力されると、文章分類部が、その文章を入力し、この入力文章を単語のデータ列と見なし、カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに対して、確率モデル記憶部に記憶される該カテゴリの線形結合モデルの該入力文章に対する負対数尤度を計算し、計算された負対数尤度の最も小さいカテゴリに該入力文章を分類する。

【００１２】また、本発明の第２の階層型文章分類装置は、ノードが文章の分類されたカテゴリを表現し、リンクがカテゴリの上位下位関係を表現するグラフとして、カテゴリの階層を記憶するカテゴリ階層記憶部、前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに対して、より下位のカテゴリの、単語空間上の確率モデルの集合を該カテゴリの確率モデルの集合とし、各カテゴリの確率モデルの集合の全ての要素を記憶する確率モデル集合記憶部、前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに分類された文章を基に、各カテゴリの確率モデルの集合を、より下位のカテゴリの単語空間上の確率モデルの集合から学習し、学習できた各カテゴリの確率モデルの集合のすべての要素を前記確率モデル集合記憶部に記憶する学習部、新しく文章を入力し、該入力文章を単語のデータ列と見なし、前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに対して、前記確率モデル集合記憶部に記憶される該カテゴリの確率モデルの集合に対する該入力文章の確率的複雑度を計算し、計算された確率的複雑度の最も小さいカテゴリに該入力文章を分類する文章分類部、を備える。

【００１３】このように構成された第２の階層型文章分類装置にあっては、学習部が、カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに例えば事前に人手によって分類された文章を基に、各カテゴリの確率モデルの集合を、より下位のカテゴリの単語空間上の確率モデルの集合から学習し、学習できた各カテゴリの確率モデルの集合のすべての要素を確率モデル集合記憶部に記憶し、その後、自動分類対象となる文章が入力されると、文章分類部が、その文章を入力し、この入力文章を単語のデータ列と見なし、カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに対して、確率モデル集合記憶部に記憶される該カテゴリの確率モデルの集合に対する該入力文章の確率的複雑度を計算し、計算された確率的複雑度の最も小さいカテゴリに該入力文章を分類する。

【００１４】

【発明の実施の形態】次に本発明の実施の形態の例について図面を参照して詳細に説明する。

【００１５】図１を参照すると、本発明の第１の実施例は、カテゴリ階層記憶部１、確率モデル記憶部２、学習部３、および文章分類部４から構成される。

【００１６】カテゴリ階層記憶部１ではカテゴリの階層が記憶される。カテゴリの階層構造はグラフとして表される。グラフでは、ノードがカテゴリを表現し、リンクがカテゴリの上位下位関係を表現する。また、カテゴリには既に分類された文章が入っている。図２にカテゴリ階層の例を示す。ここでは、カテゴリの階層が木構造となっているが、一般的にはカテゴリの階層がもっと複雑なグラフ構造になる。

【００１７】確率モデル記憶部２では、カテゴリの階層における一つのカテゴリに対して一つの確率モデルを対応させて記憶する。本実施例では、この確率モデルが線形結合モデルの形をとることを特徴とする。ある確率の線形結合モデルは、それより下位のカテゴリの確率モデルの重みつき平均として定義される。以下に線形結合モ

デルの例を示す。

【００１８】線形結合モデルの例；カテゴリの木構造では、ノードがカテゴリを表す。ノードｃのカテゴリの線形結合モデルはその子ノードのカテゴリの線形結合モデル、およびノードｃ自身に属する確率モデルの線形結合として以下のように定義される。

【数１】

$$P(W|c)=\sum_{i=1}^{n}P(W|ci)\times P(ci|c)+P(W|c')\times P(c'|c)$$

数１において、確率変数Ｗは単語の集合Ｗ＝｛ｗ１，ｗ２，…，ｗｓ｝の値をとる。Ｐ（Ｗ｜ｃ１），Ｐ（Ｗ｜ｃ２），…，Ｐ（Ｗ｜ｃｎ）はｃの子ノードｃ１，ｃ２，…，ｃｎのカテゴリの線形結合モデルである。Ｐ（Ｗ｜ｃ’）はノードｃ自身に属する確率モデルである。つまり、Ｐ（Ｗ｜ｃ’）はｃの表すカテゴリに属し、$c_1, \cdots, c_n$ の表すカテゴリに属さない確率モデルである。Ｐ（ｃ’｜ｃ），Ｐ（ｃ１｜ｃ），…，Ｐ（ｃｎ｜ｃ）はｃ’，ｃ１，…，ｃｎの事前確率である。

【００１９】学習部３は、カテゴリ階層記憶部１に記憶されるカテゴリの階層を参照し、既にカテゴリに分類された文章から各カテゴリの線形結合モデルを学習し、学習できた線形結合モデルを確率モデル記憶部２に記憶する。

【００２０】文章分類部４は、新しく文章を入力し、該文章を単語のデータ列と見なし、カテゴリ階層記憶部１に記憶されるカテゴリの階層における各カテゴリを参照し、各カテゴリに対して、確率モデル記憶部２から、そのカテゴリの対応する線形結合モデルを参照し、該文章に対する各線形結合モデルの負対数尤度を計算し、負対数尤度のもっとも小さい線形結合モデルに対応するカテゴリに該文章を分類する。

【００２１】学習部３は、幾つかの方法で線形結合モデルを学習することができる。例えば、その下位カテゴリの線形結合モデルをヒストグラムとして推定することができる。また、重み係数をＥＭアルゴリズムと呼ばれるアルゴリズムによって学習することができる。

【００２２】ここでは、学習部３の学習アルゴリズムの一例を示す。階層を表すグラフは木構造をもつとする。学習部３は、木構造となるカテゴリの階層を参照し、ボトムアップ的にカテゴリの線形結合モデルを学習する。その学習アルゴリズムは以下の通りであり、そのフローチャートを図３に示す。

【００２３】

ノードｃを入力とする。最初は、木構造のルートノードが入力される。
ｉｆ
　ノードｃは葉ノードである。
ｔｈｅｎ
　ノードｃのカテゴリに分類された30.章から、ｃの線形結合モデルを学習し、戻る。
ｅｌｓｅ
　ノードｃの子ノードｃｉ（ｉ＝１，２，…，ｎ）の線形結合モデルを参照する。
　ｉｆ
　　ノードｃｉの線形結合モデルはまだ学習できていない。
　ｔｈｅｎ
　　ノードｃｉに対して、再帰的に本アルゴリズムを適用する。
　ｅｌｓｅ
　　ノードｃｉの線形結合モデルとｃ自身の確率モデルからノードｃの線形結合モデルを学習し、戻る。

【００２４】文章分類部４は文章の統計的仮説検定によって文章を分類する。次に、文章分類部４のアルゴリズムの一例を示し、そのフローチャートを図４に示す。

【００２５】

ｄは入力された文章であるとする。ノードｃと文章ｄを入力とする。最初は、木構造のルートノードが入力される。
ｉｆ
　ノードｃは葉ノードである。
ｔｈｅｎ
　文章ｄはノードｃのカテゴリに属するとし、終了する。
ｅｌｓｅ

文章ｄに対するノードｃの線形結合モデルの負対数尤度Ｌ（ｄ｜ｃ）を計算する。ノードｃの子ノードｃｉ（ｉ＝１，２，…，ｎ）の負対数尤度Ｌ（ｄ｜ｃｉ）をも計算する。計算できたＬ（ｄ｜ｃ）とＬ（ｄ｜ｃｉ）の最小値を求める。

ｉｆ
　子ノードの中のｃｉの負対数尤度が最小である。
ｔｈｅｎ
　ノードｃｉに対して本アルゴリズムを再帰的に適用する。
ｅｌｓｅ
　文章ｄはノードｃのカテゴリに属するとし、終了する。

【００２６】次に、学習部３による線形結合モデルを学習する方法と、文章分類部４による負対数尤度の計算方法を、さらに具体的な例を通じて説明する。カテゴリの階層は図５に示すものとする。図５中、ｃ１，ｃ２，ｃ３はカテゴリであり、ｄ１，ｄ２，ｄ３は既に分類された文章である。また、図６に各文章ｄ１，ｄ２，ｄ３における単語ｗ１，ｗ２，ｗ３の出現頻度を示す。単語ｗ１，ｗ２，ｗ３は予め定められたキーワードである。

【００２７】○線形結合モデルの学習の例
ｃ２とｃ３は葉ノードであるので、それらのノードのカテゴリの線形結合モデルは文章における単語のヒストグラムとして、図７（ａ）のように学習される。

【００２８】ｃ１に分類された文章ｄ２から、ｃ１自身に属する確率モデルを単語のヒストグラムとして学習する。これをＰ（Ｗ｜ｃ１’）と表す。つまり、それはｃ１のカテゴリに属し、ｃ２，ｃ３のカテゴリに属さない確率モデルであり、図７（ｂ）のように学習される。

【００２９】一方、各モデルの事前分布を以下のように学習する。

【数２】

$$P(ci \mid c) = \frac{f\ (ci)}{N}$$

【００３０】ここで、ｆ（ｃｉ）はノードｃｉとその支配するノードの属する文章数で、Ｎは全文章数である。よって、各モデルの事前分布は図７（ｃ）のように学習される。

【００３１】さらに、線形結合モデルの定義に従って、ノードｃ１における線形結合モデルを以下のように学習することができる。

【数３】

$$P(w1 \mid c1) = P(w1 \mid c1') \times P(c1' \mid c1) + P(w1 \mid c2) \times P(c2 \mid c1) + P(w1 \mid c3) \times P(c3 \mid c1)$$
$$= \frac{1}{2} \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{3} = \frac{4}{9}$$

$$P(w2 \mid c1) = P(w2 \mid c1') \times P(c1' \mid c1) + P(w2 \mid c2) \times P(c2 \mid c1) + P(w2 \mid c3) \times P(c3 \mid c1)$$
$$= \frac{1}{4} \times \frac{1}{3} + \frac{1}{6} \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{3} = \frac{11}{36}$$

$$P(w3 \mid c1) = P(w3 \mid c1') \times P(c1' \mid c1) + P(w3 \mid c2) \times P(c2 \mid c1) + P(w3 \mid c3) \times P(c3 \mid c1)$$
$$= \frac{1}{4} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{3} + \frac{1}{6} \times \frac{1}{3} = \frac{1}{4}$$

【００３２】即ち、ノードｃ１における線形結合モデルは図７（ｄ）に示すようになる。

【００３３】○負対数尤度の計算の例
新しい文章ｄにおける単語の分布は図８に示すものとする。つまり、文章分類部４は入力文章中から単語ｗ１を２個、単語ｗ２を１個、単語ｗ３を１個検出したとする。ｄに対するｃ１の負対数尤度を以下のように計算する。対数の底は２であるとする。

【数４】

$$L(d \mid c1) = 2 \cdot -\log \frac{4}{9} - \log \frac{11}{36} - \log \frac{1}{4}$$
$$= 6.05$$

【００３４】同様に、ｃ２，ｃ３の負対数尤度を計算する。

【数５】

$$L(d|c2) = 2 \cdot -\log\frac{1}{2} - \log\frac{1}{6} - \log\frac{1}{3}$$

$$= 6.17$$

【数６】

$$L(d|c3) = 2 \cdot -\log\frac{1}{3} - \log\frac{1}{2} - \log\frac{1}{6}$$

$$= 6.75$$

【００３５】尤度Ｌ（ｄ｜ｃ１）がもっとも小さいので、ｄはｃ１に分類される。

【００３６】図９を参照すると、本発明の第２の実施例は、カテゴリ階層記憶部１、確率モデル集合記憶部５、学習部６、および文章分類部７から構成される。

【００３７】カテゴリ階層記憶部１ではカテゴリの階層が記憶される。カテゴリの階層では、ノードがカテゴリを表し、リンクが上位下位関係を表す。カテゴリ階層の例として前述した図２がある。

【００３８】確率モデル集合記憶部５では、確率モデルの集合が記憶される。カテゴリの階層における各カテゴリに対して一つの確率モデルの集合が定義され、記憶される。以下に確率モデルの集合の例を示す。

【００３９】〇確率モデル集合の例

ノードｃの確率モデルの集合が確率モデルＰ（Ｗ｜ｃ’），Ｐ（Ｗ｜ｃ１），…，Ｐ（Ｗ｜ｃｎ）を含むとする。Ｐ（Ｗ｜ｃ１），…，Ｐ（Ｗ｜ｃｎ）はｃの子ノード$c_1, \cdots c_n$ の確率モデルの集合のもつ確率モデル（確率分布）である。Ｐ（Ｗ｜ｃ’）はノードｃ自身に属する確率モデルである。つまり、それは、ｃのカテゴリに属し、ｃ１，…，ｃｎのカテゴリに属さない確率モデルである。また、各確率モデルの事前確率Ｐ（ｃ’｜ｃ），Ｐ（ｃ１｜ｃ），…，Ｐ（ｃｎ｜ｃ）が存在するとする。確率モデルＰ（Ｗ｜ｃ’），Ｐ（Ｗ｜ｃ１），…，Ｐ（Ｗ｜ｃｎ）は、例えば、ヒストグラムの形で表現される。

【００４０】各カテゴリの確率モデルの集合は、それ自身に属する文章による単語空間上の確率モデルと、その下位のカテゴリに属する文章による単語空間上の確率モデルからなる。

【００４１】学習部６は、カテゴリ階層記憶部１に記憶されるカテゴリの階層を参照し、既にカテゴリに分類された文章から各カテゴリの対応するモデル集合を学習し、学習できた確率モデルの集合を確率モデル集合記憶部５に記憶する。

【００４２】文章分類部７は、新しく文章を入力し、該文章を単語のデータ列と見なし、カテゴリ階層記憶部１に記憶されるカテゴリにおける各カテゴリを参照し、各カテゴリに対して、確率モデル集合記憶部５から、そのカテゴリの対応する確率モデル集合を参照し、該文章の各参照された確率モデル集合に対する確率的複雑度を計算し、確率的複雑度のもっとも小さい確率モデル集合に対応するカテゴリに該文章を分類する。

【００４３】確率的複雑度とは、確率モデルの集合を用いてデータを記述する際の最小記述長を表す量で、リッサネン（Ｒｉｓｓａｎｅｎ）によって提唱されたものである（Ｊｏｒｍａ Ｒｉｓｓａｎｅｎ，Ｓｔｏｃｈａｓｔｉｃ Ｃｏｍｐｌｅｘｉｔｙ ｉｎ Ｓｔａｔｉｓｔｉｃａｌ Ｉｎｑｕｉｒｙ，Ｗｏｒｌｄ Ｓｃｉｅｎｔｉｆｉｃ Ｐｕｂｌｉｓｈｉｎｇ Ｃｏ．，Ｓｉｎｇａｐｏｒｅ，１９８９）。本実施例では、確率的複雑度を、確率モデル集合における確率モデルのデータに対する尤度の重み付き平均の負対数として計算する。

【００４４】次に、学習部６の学習アルゴリズムの一例を示す。階層を表すグラフが木構造をもつとする。学習部６は、木構造となるカテゴリの階層を参照し、ボトムアップ的にカテゴリの確率モデル集合を学習する。そのアルゴリズムは以下の通りであり、そのフローチャートを図１０に示す。

【００４５】

ノードｃを入力とする。最初は、木構造のルートノードが入力される。

ｉｆ

　ノードｃは葉ノードである。

ｔｈｅｎ

　ノードｃのカテゴリに分類された文章から、ｃの確率モデル集合の全ての要素を学習し、戻る。

ｅｌｓｅ

　ノードｃの子ノードｃｉ（ｉ＝１，２，…，ｎ）の確率モデル集合を参照する。

　ｉｆ

　　ノードｃｉの確率モデル集合はまだ学習できていない。

　ｔｈｅｎ

ノードｃｉに対して、再帰的に本アルゴリズムを適用する。

ｅｌｓｅ

ノードｃｉの確率モデル集合とｃに分類された文章の確率モデルからノードｃの確率モデル集合を学習し、戻る。

【００４６】文章分類部７は統計的仮説検定によって文章を分類する。次に、文章分類部７のアルゴリズムの一例を示す。図１１はそのフローチャートである。

【００４７】

ｄは入力された文章であるとする。ノードｃと文章ｄを入力とする。最初は、木構造のルートノードが入力される。

ｉｆ

ノードｃは葉ノードである。

ｔｈｅｎ

文章ｄはノードｃのカテゴリに属するとし、終了する。

ｅｌｓｅ

ノードｃにおける文章ｄの確率的複雑度ＳＣ（ｄ｜ｃ）を計算する。ノードｃの子ノードｃｉ（ｉ＝１，２，…，ｎ）における確率的複雑度ＳＣ（ｄ｜ｃｉ）をも計算する。計算できたＳＣ（ｄ｜ｃ）とＳＣ（ｄ｜ｃｉ）の中の最小値を求める。

ｉｆ

ノードの中のｃｉの確率的複雑度が最小である。

ｔｈｅｎ

ノードｃｉに対して本アルゴリズムを再帰的に適用する。

ｅｌｓｅ

文章ｄはノードｃのカテゴリに属するとし、終了する。

【００４８】次に確率的複雑度の計算例を示す。

【００４９】カテゴリの階層は図５に示すものとする。また、文章における単語（キーワード）の出現頻度は図６に示すものであるとする。

【００５０】ノードｃ２，ｃ３が葉ノードであるので、それぞれのもつ確率モデルの集合は一つの確率モデルを含む。さらに、それらの確率モデルがヒストグラムとして、図１２（ａ）のように学習される。

【００５１】ノードｃ１自身に属する確率モデルもヒストグラムとして、図１２（ｂ）のように学習される。

【００５２】従って、ｃ１の確率モデル集合は確率モデルＰ（Ｗ｜ｃ１），Ｐ（Ｗ｜ｃ２），Ｐ（Ｗ｜ｃ３）を含むことになる。それらの確率モデルの事前確率Ｐ（ｄｉ｜ｃ）が一様分布であるとする。

【００５３】新しい文章ｄにおける単語の出現頻度は図１２（ｃ）に示すものであるとする。

【００５４】ｄのｃ１に対する確率的複雑度を以下のように計算する。対数の底は２であるとする。

【数７】

$$SC(d|c1) = -\log\left(\frac{1}{3}\cdot\left(\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{6}\cdot\frac{1}{3}+\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{4}\cdot\frac{1}{4}+\frac{1}{3}\cdot\frac{1}{3}\cdot\frac{1}{2}\cdot\frac{1}{6}\right)\right)$$

【００５５】ｄのｃ２，ｃ３に対する確率的複雑度を以下のように計算する。

【数８】

$$SC(d|c2) = -\log\left(\frac{1}{3}\cdot\frac{1}{3}\cdot\frac{1}{2}\cdot\frac{1}{6}\right) = 6.75$$

【数９】

$$SC(d|c3) = -\log\left(\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{6}\cdot\frac{1}{3}\right) = 6.17$$

【００５６】ＳＣ（ｄ｜ｃ３）がもっとも小さいので、ｄはｃ３に分類される。

【００５７】図１３は本発明の階層型文章分類装置の第３の実施例のブロック図である。この例の階層型文章分類装置は、ＣＰＵ１０１、主記憶１０２および補助記憶１０３を含むコンピュータ１０４と、このコンピュータ１０４に接続された表示装置１０５、入力装置１０６およびファイル１０７を含む入出力装置１０８と、階層型文章分類プログラムを記録する記録媒体１０９とから構成される。記録媒体１０９はＣＤ－ＲＯＭ、半導体メモリ等の機械読み取り可能な記録媒体であり、ここに記録された階層型文章分類プログラムは、コンピュータ１０４に読み取られ、コンピュータ１０４の動作を制御することにより、コンピュータ１０４上に、図１に示したカテゴリ階層記憶部１、確率モデル記憶部２、学習部３および文章分類部４、または図９に示したカテゴリ階層記憶部１、確率モデル集合記憶部５、学習部６および文章分類部７を実現する。

【００５８】

【発明の効果】以上説明したように、本発明によれば、

階層構造をなすカテゴリに文章を自動分類することができ、かつ尤度比検定の理論に基づいた統計的信頼性の高い文章分類ができる。

【図面の簡単な説明】

【図１】本発明の階層型文章分類装置の第１の実施例のブロック図である。

【図２】カテゴリ階層の例を示す図である。

【図３】本発明の階層型文章分類装置の第１の実施例における学習アルゴリズムの一例を示すフローチャートである。

【図４】本発明の階層型文章分類装置の第１の実施例における文章分類のアルゴリズムの一例を示すフローチャートである。

【図５】カテゴリ階層の例を示す図である。

【図６】文章における単語分布の例を示す図である。

【図７】線形結合モデルの学習例の説明図である。

【図８】負対数尤度の計算例の説明図である。

【図９】本発明の階層型文章分類装置の第２の実施例のブロック図である。

【図１０】本発明の階層型文章分類装置の第２の実施例における学習アルゴリズムの一例を示すフローチャートである。

【図１１】本発明の階層型文章分類装置の第２の実施例における文章分類のアルゴリズムの一例を示すフローチャートである。

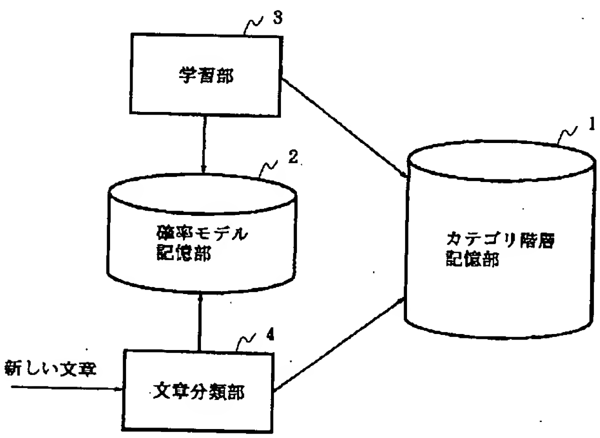
【図１２】確率的複雑度の計算例の説明図である。

【図１３】本発明の階層型文章分類装置の第３の実施例のブロック図である。

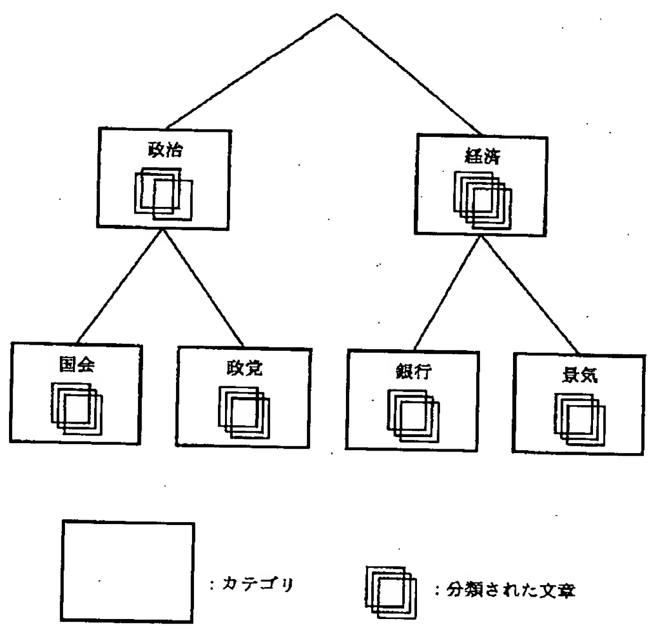
【符号の説明】

1　カテゴリ階層記憶部
2　確率モデル記憶部
3　学習部
4　文章分類部
5　確率モデル集合記憶部
6　学習部
7　文章分類部

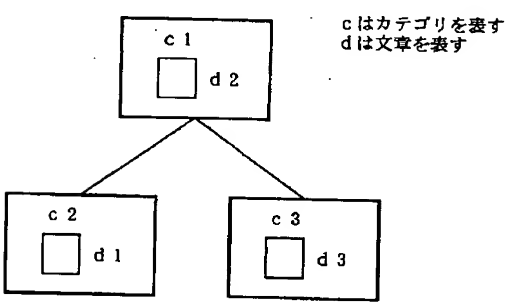
【図１】



【図２】



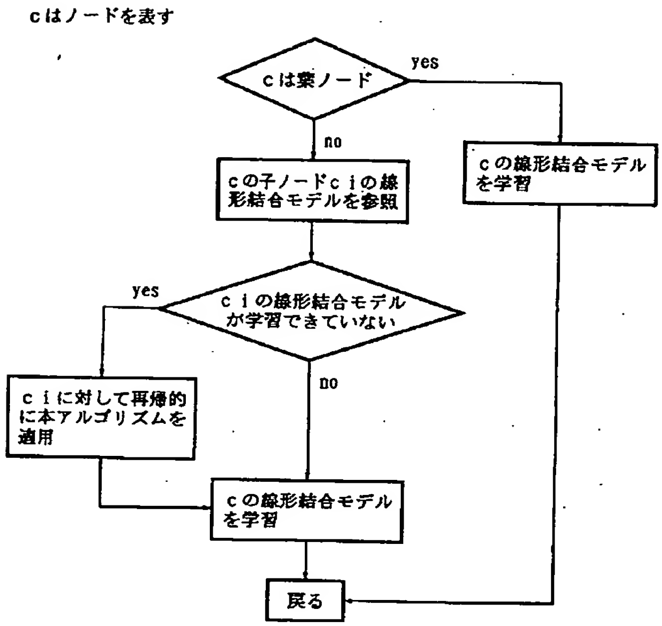
【図５】



【図８】

	w 1	w 2	w 3
f (W)	2	1	1

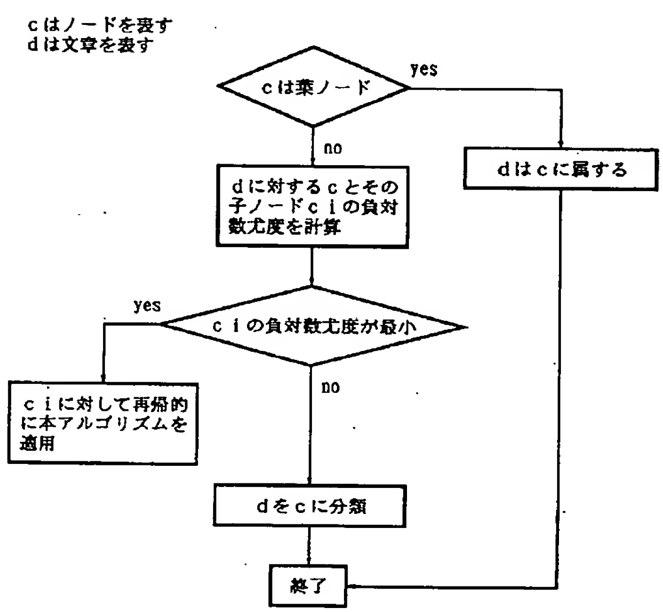
【図3】



【図6】

f（W｜D）	w1	w2	w3
d1	2	0	1
d2	1	0	0
d3	1	2	0

【図4】



【図7】

（a）

	w1	w2	w3
P（W｜c2）	1/2	1/6	1/3
P（W｜c3）	1/3	1/2	1/6

（b）

	w1	w2	w3
P（W｜c1'）	1/2	1/4	1/4

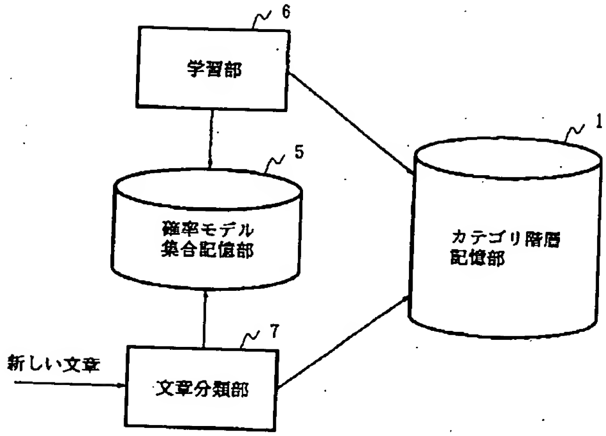
（c）

P（c2｜c1）＝1/3
P（c3｜c1）＝1/3
P（c1'｜c1）＝1/3

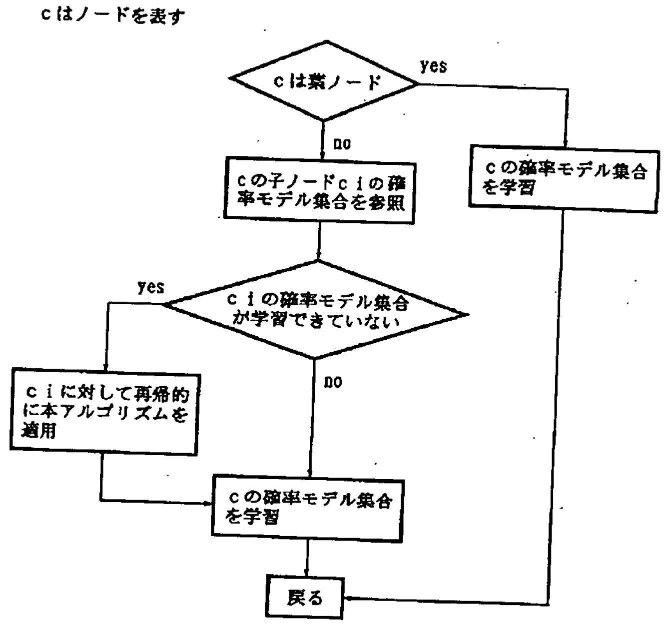
（d）

	w1	w2	w3
P（W｜c1）	4/9	11/36	1/4

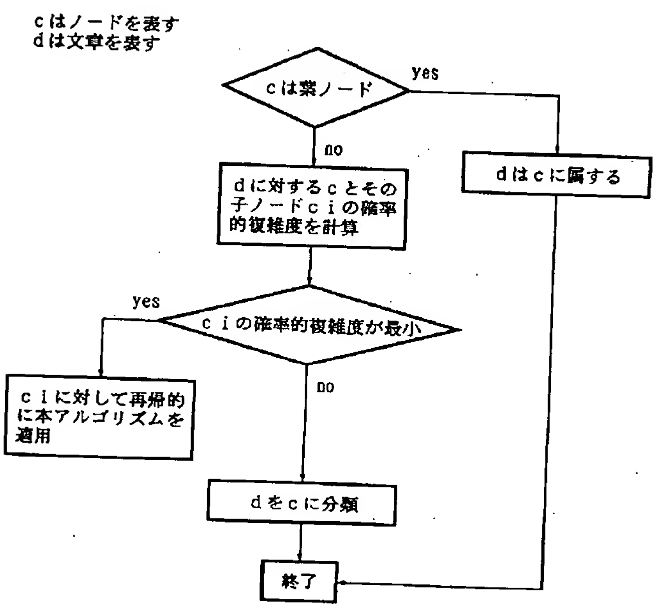
【図９】



【図１０】



【図１１】



【図１２】

(a)

	w 1	w 2	w 3
P (W | c2)	1/2	1/6	1/3
P (W | c3)	1/3	1/2	1/6

(b)

	w 1	w 2	w 3
P (W | c3)	1/2	1/4	1/4

(c)

	w 1	w 2	w 3
f (W | d3)	2	1	1

【図１３】